Analysis of Climate Data in Response to Soil Temperature and Moisture

INTRODUCTION

This analysis is about the application of statistical methods to climate data. The data relates to understanding soil and air temperatures (as variables of interest) based on different factors. These factors are represented by 9 variables associated to relative humidity, dew point, and soil moisture. Our goal is to identify the relationship between these factors and our variables of interest and understand the impact of such factors on the variables of interest.

Data is sourced from weather sensors that collect and store the information every five minutes, 24 hours a day. The sample taken is from February 15th to March 30th, 2024. The sensors are on a tree and a light post in the FIU MMC campus next to the Ocean Bank Arena building.

Given the goal above, we set the following objectives:

- Develop a model to appreciate the daily impact of relative humidity and dew point to soil temperature. This is called "Model 1" in the study.
- Develop a model to appreciate the impact of relative humidity and dew point to soil moisture in the midnight and midday. This will enable us to make a comparison of said impact which is expected to be different during the day and during the night. Two models resulted: "Model 2" for the results at noon time and "Model 3" for the results at midnight.
- Develop a model to appreciate the impact of relative humidity and dew point to soil temperature. This model is called "Model 4".

DATA EXPLANATION

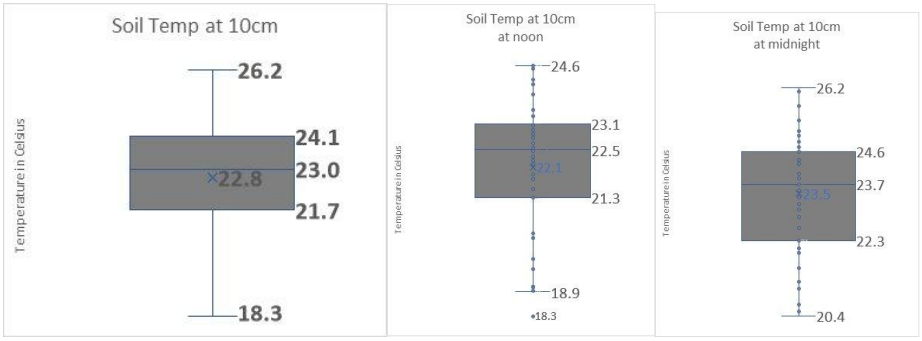
The date and time when each record is logged were used to identify midday and midnight. These records were logged every 5 minutes. To simplify our study, the records were summarized by averaging the values of the following predictors:

- air temperature, relative humidity, and dew point as captured from the sensors placed at 6 feet of height on a tree;
- air temperature, relative humidity, and dew point as captured from the sensors placed at 6 feet of height on a light post;
- soil moisture as captured from the sensors placed 15 centimeters under the ground;
- and soil temperature from the sensors placed 10 centimeters of ground depth.

In total, we obtained readings for 9 variables, 2 of which separately represent the outcome variables and the remaining 7 the predictors. For the nearly 45-day period covered, we obtained the following number of records as our samples per model:

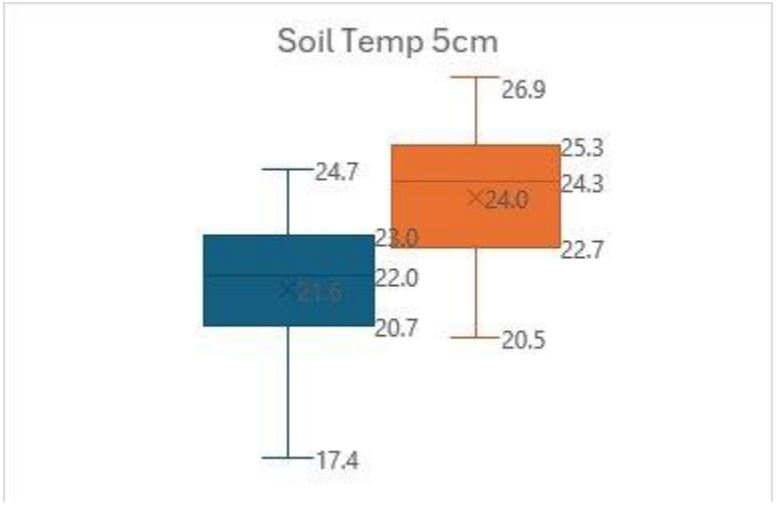
- 43records for Model 1
- 86 records (43 for the midday and 43 for the midnight) for Models 2 and 3, respectively.
- 43 records for Model 4

We notice an interesting behavior of the soil temperature during the highest point of the day (noon) and the lowest point of the day (midnight). The soil takes time to warm up during the day and it preserves the warmth until late night and cools off until midday. The graphs below depict the behavior quantitatively.



To confirm this phenomenon, we summarized the climate readings of the same period in different time intervals to confirm our understanding. Looking at the soil temperature at only 5cm of depth, we found that, during the same 45-day period, the soil temperature diminishes in the morning and increases in the night. This is better explained by the graph below.

Morning(8-9AM) - Blue
Evening(8-9PM)-Orange



METHODS APPLIED

All models are constructed using soil moisture as the dependent variable, with various climate-related predictors such as dew point and humidity from multiple locations. Initially, this model was evaluated using multiple linear regression, followed by verification of the assumptions necessary for linear regression. We also conducted multicollinearity tests and later applied Lasso and Ridge regression techniques and Principal Component Regression.

Multiple Linear Regression

This regression method fits a model to show the linear relationship between dependent variable (Soil Temperature for Model1, Model2, Model 3 and Soil Moisture for Model 4) and independent variables such as Air Temperature, Relative humidity, Dew point

$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + €$

- Y is the dependent variable.
- β0 is the intercept.
- β1, β2,βn are the coefficients of the predictors X1, X2 etc.
- € is the error term, assumed to be normally distributed.

Regularized Regression

We fitted the model for different data by cross-validation, then selected the minimum lambda, which is a parameter to control the shrinkage effect on the coefficients of predictors and reduces prediction error. ++With that minimum lambda, we again fitted the model and obtained the coefficients for both Ridge and Lasso regression. Here, the Ridge tries to fit the model well and reduce complexity whereas the lasso tries to reduce complexity.

Principal Component Regression and Partial Least Squares

To avoid the multicollinearity and high dimensionality of variables from multiple locations of our data, PCR converts the predictors into a set of principal components to capture the variance of target variables. When it comes to PLS, it not only reduces the number of dimensions but also captures the variance in soil temperature for Models 1, 2, 3 and soil moisture for Model 4. It helps us to handle the complexities of our environment data and gives accurate results for our target variables according to model respectively.

RESULTS AND DISCUSSION

Model 1

Multiple Linear Regression (Target variable – SoilTemp_10cm)

SoilTemp_10cm = 8.73(AirTemp_6ft_CP) + 1.34(RH_6ft_CP) - 11.3(DP_6ft_CP) + 1.57(AirTemp_6ft_T1) + 0.73(RH_6ft_T1) + 5.05(DP_6ft_T1) - 8.43(AirTemp_6ft_T2) - 1.75(RH_6ft_T2) + 4.88(DP_6ft_T2) + €

- For each unit increase in AirTemp_6ft_CP, the Soil_Temp_10cm is increased by 8.73 units while holding all other independent variables constant.
- For each unit increase in RH_6ft_CP, the Soil_Temp_10cm is increased by 1.34 units while holding all other independent variables constant.
- For each unit increase in DP_6ft_CP, the Soil_Temp_10cm is decreased by 11.3 units while holding all other independent variables constant.
- For each unit increase in AirTemp_6ft_T1, the Soil_Temp_10cm is increased by 1.57 units while holding all other independent variables constant.
- For each unit increase in RH_6ft_T1, the Soil_Temp_10cm is increased by 0.73 units while holding all other independent variables constant.
- For each unit increase in DP_6ft_T1, the Soil_Temp_10cm is increased by 5.05 units while holding all other independent variables constant.
- For each unit increase in AirTemp_6ft_T2, the Soil_Temp_10cm is decreased by 8.43 units while holding all other independent variables constant.
- For each unit increase in RH_6ft_T2, the Soil_Temp_10cm is decreased by 8.73 units while holding all other independent variables constant.
- For each unit increase in DP_6ft_T2, the Soil_Temp_10cm is increased by 4.88 units while holding all other independent variables constant.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -19.1994    20.7572  -0.925    0.362
AirTemp_6ft_CP   8.7382    10.3836   0.842    0.406
RH_6ft_CP        1.3406     2.2877   0.586    0.562
DP_6ft_CP      -11.3037    11.2467  -1.005    0.322
AirTemp_6ft_T1   1.5798    11.5314   0.137    0.892
RH_6ft_T1        0.7396     2.5459   0.291    0.773
DP_6ft_T1        5.0516    11.4987   0.439    0.663
AirTemp_6ft_T2  -8.4346     9.3090  -0.906    0.371
RH_6ft_T2       -1.7586     1.9501  -0.902    0.374
DP_6ft_T2        4.8835     9.0206   0.541    0.592

Residual standard error: 0.7026 on 33 degrees of freedom
Multiple R-squared:  0.8584,    Adjusted R-squared:  0.8197
F-statistic: 22.22 on 9 and 33 DF,  p-value: 1.459e-11
```

T-test for Regression Coefficient's:
Hypothesis Testing
Null Hypothesis: The coefficient of independent variable is equal to zero.
Alternate Hypothesis: The coefficient of independent variable is not equal to zero.

Here, all independent variable p-values are greater than zero. We failed to reject null hypothesis. We are 95% confident that there is no strong evidence to conclude that the predictor variable is constant.
None of the predictor variables is significant. There may be many **predictors** and **multicollinearity**. So, we can perform Regularized Regression and Principal Component analysis.

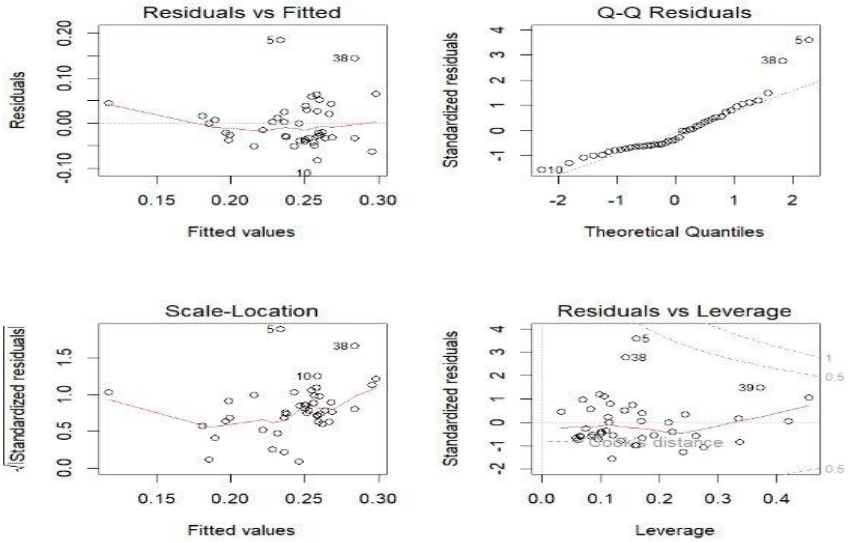
Let's see overall significance by performing Anova F-test.

Anova F-test for overall significance:
Hypothesis Testing
Null Hypothesis: All Regression coefficients are equal to zero.
Alternate Hypothesis: At least one of the regression coefficients is not zero.

Since the p-value = 1.459e-11 which is less than 0.05, we reject the null hypothesis. Then we are 95% confident and have evidence to conclude from the sampled data that at least one of the regression coefficients is not zero. And we can also conclude that the model which we use is statistically significant, overall.

81.97 percent of variability in SoilTemp_10cm is explained by all predictors in the regression model.

Assumptions



Linearity: The Residuals vs Fitted graph shows randomness of points and no trend patterns found in Graph. This indicated the Linearity assumption is valid.
Normality: The normal Q-Q plot shows that the points fall along the straight line and there is three or four points are somewhat far from line in both models. But due to the small dataset, this also indicated that Normality assumption is valid.
Homoscedasticity: It looks like residuals have constant variance against the fitted values in both models. This shows that the Homoscedasticity is also good.
Outliers: There are no points with high leverage and high residuals which indicates a good sign that there are no influential points. There are no other points deviating from horizontal line. This indicates that there are no outliers.

Multicollinearity

```
AirTemp_6ft_CP      RH_6ft_CP      DP_6ft_CP AirTemp_6ft_T1      RH_6ft_T1
      76350.65       32384.13      158790.06       87087.75       40210.26
    DP_6ft_T1 AirTemp_6ft_T2      RH_6ft_T2      DP_6ft_T2
    163467.37       60030.69       23136.03      102597.82
```

All the predictors have Variance Inflation Factor values greater than 10 and very high values. When the VIF values are greater than 10, there are highly correlated predictor variables in model or multicollinearity. Let's perform Regularized Regression.

Regularized Regression (Ridge and Lasso Regression)

	s0
(Intercept)	13.899655174
AirTemp_6ft_CP	0.133671706
RH_6ft_CP	-0.014454360
DP_6ft_CP	0.052320053
AirTemp_6ft_T1	0.109519406
RH_6ft_T1	0.001639055
DP_6ft_T1	0.050200500
AirTemp_6ft_T2	0.112437025
RH_6ft_T2	-0.006378562
DP_6ft_T2	0.047143058

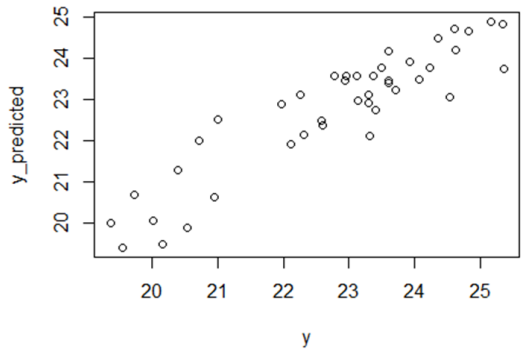
	s0
(Intercept)	10.93166066
AirTemp_6ft_CP	0.50692925
RH_6ft_CP	.
DP_6ft_CP	.
AirTemp_6ft_T1	.
RH_6ft_T1	0.01023286
DP_6ft_T1	.
AirTemp_6ft_T2	.
RH_6ft_T2	.
DP_6ft_T2	.

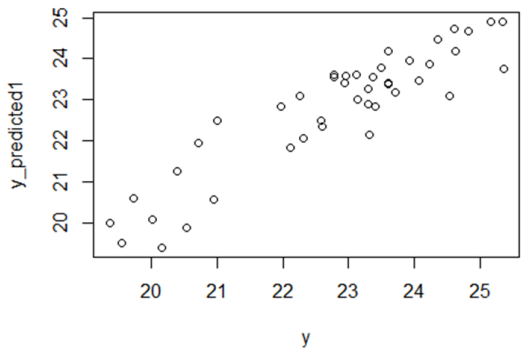
For Ridge Regression, the important predictors are AirTemp_6ft_CP, AirTemp_6ft_T1 and AirTemp_6ft_T2 because they are not close to zero and other predictors are close to zero.

For Lasso Regression the important predictors are AirTemp_6ft_CP and AirTemp_6ft_T1. It eliminated all other predictors by making its coefficients equal to zero and other predictors close to zero.

The Ridge Regression adds penalty equal to square of magnitude of coefficient to regularize the model. It shrinks the coefficients close to zero, but it does not make coefficients equal to zero. It achieved an Rsquare of 83.04 which shows it as a better model.

The Lasso Regression adds penalty equal to absolute value of magnitude of coefficient to regularize the model. It makes coefficients equal to zero and performs feature selection. It achieved an R-square 83.56 which shows it as a better model.



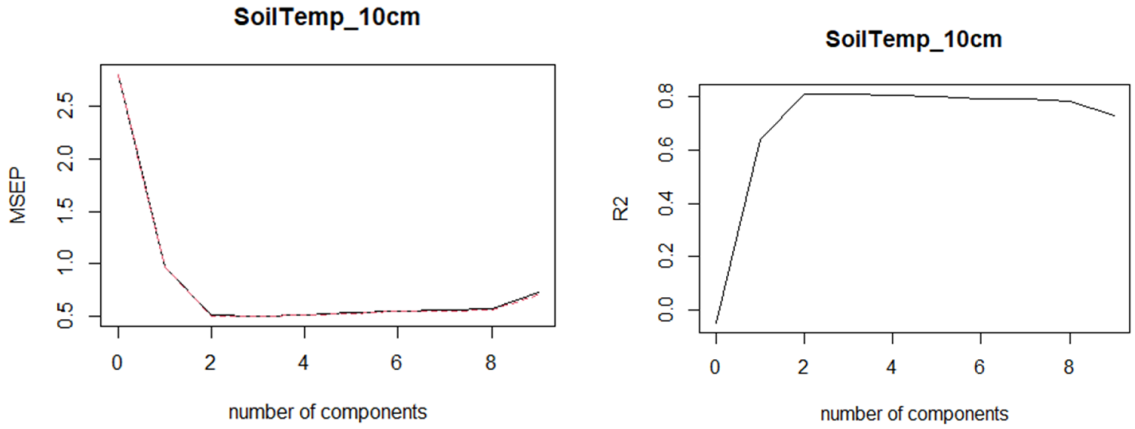


When we consider identifying the most significant predictors, we may decide the Lasso regression is a better fit for the data because it performs feature selection by eliminating some of the predictors. When we consider explaining maximum variability in response variable, we may decide that the Ridge model is a better fit for data.

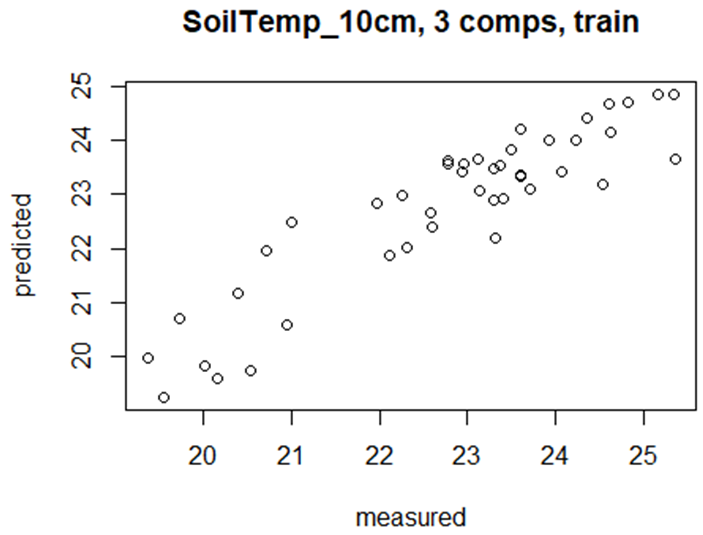
Principal Component Analysis

Principal Component Regression

Beyond the 3rd component, the Root Mean Squared Error of Prediction (RMSEP) does not exhibit appreciable improvement. In fact, the RMSEP has not decreased that much from the 3rd component. It already captured about 83.43 of Soil Temperature. Much of the relevant variation has been considered because the incremental gain in explained variance beyond the third component is negligible for both predictors and dependent variables. Overall, after the 6th components there is not much difference in RMSEP, variability explained by components on predictors and dependent variables.



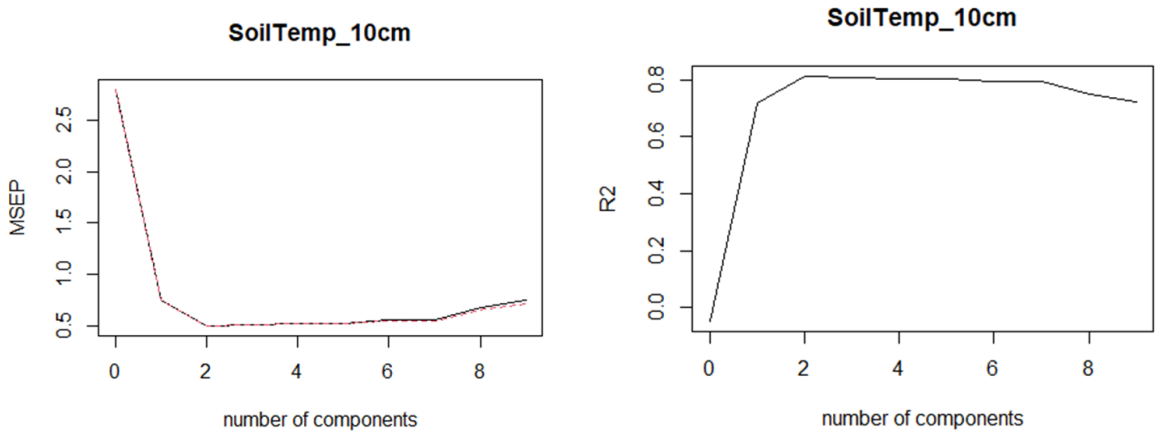
The plot between actual value of Soil Temperature and predicted values shows a trend line called regression line indicating that model can capture the trend in data. It indicates that this model is best for predicted dependent variable with our independent variables.



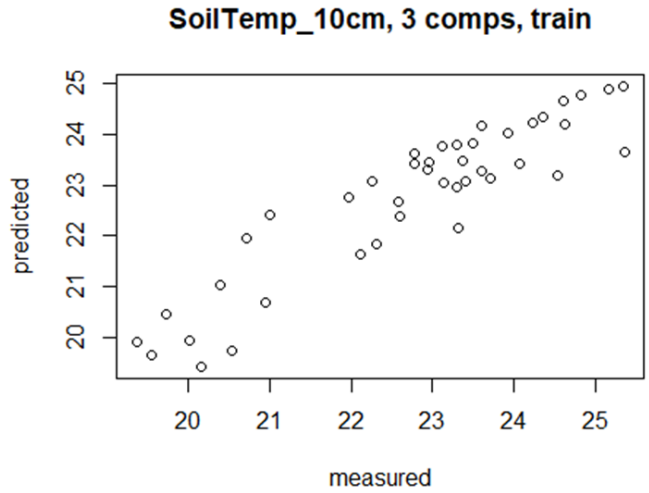
Partial Least Squares

The Root Mean Squared Error of Prediction (RMSEP) shows no discernible improvement past the 3rd component. The RMSEP really decreased a little from the 3rd component. By the third

component, it had already roughly 84.31 percent of Soil Temperature. Since the incremental improvement in explained variance beyond the third component is minimal for both predictors and dependent variables, it seems that much of the important variation has already been included.



Overall, RMSEP does not significantly differ after the third component; the variability is explained by components related to predictors and dependent variables. The plot between actual values and predicted values shows a trend line called regression line indicating that model with 3 components can capture the trend in data. It indicates that this model is best for predicted dependent variable with our independent variables.



Model 2
Multiple Linear Regression

The Model 2 is described by the following equation:

SoilTemp_10cm = 1.710 + 9.217(AirTemp_6ft_CP) + 1.776(RH_6ft_CP) - 6.364(DP_6ft_CP) + 6.250(AirTemp_6ft_T1) + 2.667(RH_6ft_T1) - 5.809(DP_6ft_T1) - 14.519(AirTemp_6ft_T2) - 4.340(RH_6ft_T2) + 11.774(DP_6ft_T2).

That is, during the midnight hour (from 12am to 1am), the average temperature of the soil at 10 cm of depth expressed in Celsius can be described by the averaged Air Temperature, the Relative Humidity, and the Dew Point at 6 feet height on the post, at 6 feet height of the Tree 1, and at the same height on Tree 2, respectively.

The specific interpretation of the coefficients of the model is as follows… each coefficient represents the change in soil temperature associated with a one-unit change in the predictor variable, holding other variables constant:

- The soil temperature would start at 1.710 degrees Celsius if none of the predictors were present.
- AirTemp_6ft_CP: A one-unit increase in the average air temperature at 6 feet in location CP (the post) increases soil temperature by about 9.217 units.
- RH_6ft_CP: A one-unit increase in the average relative humidity at 6 feet in location CP (the post) increases soil temperature by about 0.1.776 units.
- DP_6ft_CP: A one-unit increase in the average dew point at 6 feet in location CP (the post) decreases soil temperature by about 6.364 units.
- AirTemp_6ft_T1: A one-unit increase in the average air temperature at 6 feet in location T1 (three 1) increases soil temperature by about 6.250 units.
- RH_6ft_T1: A one-unit increase in the average relative humidity at 6 feet in location T1 (tree 1) increases soil temperature by about 2.667 units.
- DP_6f_T1: A one-unit increase in the average dew point at 6 feet in location T1 (tree 1) decreases soil temperature by about 5.809 units.
- AirTemp_6ft_T2: A one-unit increase in the average air temperature at 6 feet in location T1 (tree 1) decreases soil temperature by about 14.519 units.
- RH_6ft_T2: A one-unit increase in the average relative humidity at 6 feet in location T2 (tree 2) decreases soil temperature by about 4.340 units.
- DP_6ft_T2: A one-unit increase in the average dew point at 6 feet in location T2 (tree 2) increases by about 11.774 unit

```
Call:
lm(formula = SoilTemp_10cm ~ ., data = midnight_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.21611 -0.58192  0.08437  0.39162  1.58204

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.710     14.999   0.114    0.910
AirTemp_6ft_CP   9.217     18.591   0.496    0.623
RH_6ft_CP        1.776      4.175   0.425    0.673
DP_6ft_CP       -6.364     18.832  -0.338    0.738
AirTemp_6ft_T1   6.250     12.228   0.511    0.613
RH_6ft_T1        2.667      3.143   0.849    0.402
DP_6ft_T1       -5.809     16.243  -0.358    0.723
AirTemp_6ft_T2 -14.519     12.347  -1.176    0.248
RH_6ft_T2       -4.340      2.820  -1.539    0.133
DP_6ft_T2       11.774      8.270   1.424    0.164

Residual standard error: 0.7098 on 33 degrees of freedom
Multiple R-squared:  0.8396,    Adjusted R-squared:  0.7959
F-statistic:  19.2 on 9 and 33 DF,  p-value: 1.054e-10
```

T-test for Regression Coefficient's:
Hypothesis Testing
Null Hypothesis: The coefficient of independent variable is equal to zero.
Alternate Hypothesis: The coefficient of independent variable is not equal to zero.

Here, all independent variable p-values are greater than zero. We failed to reject null hypothesis. We are 95% confident that there is no strong evidence to conclude that the predictor variable is constant.

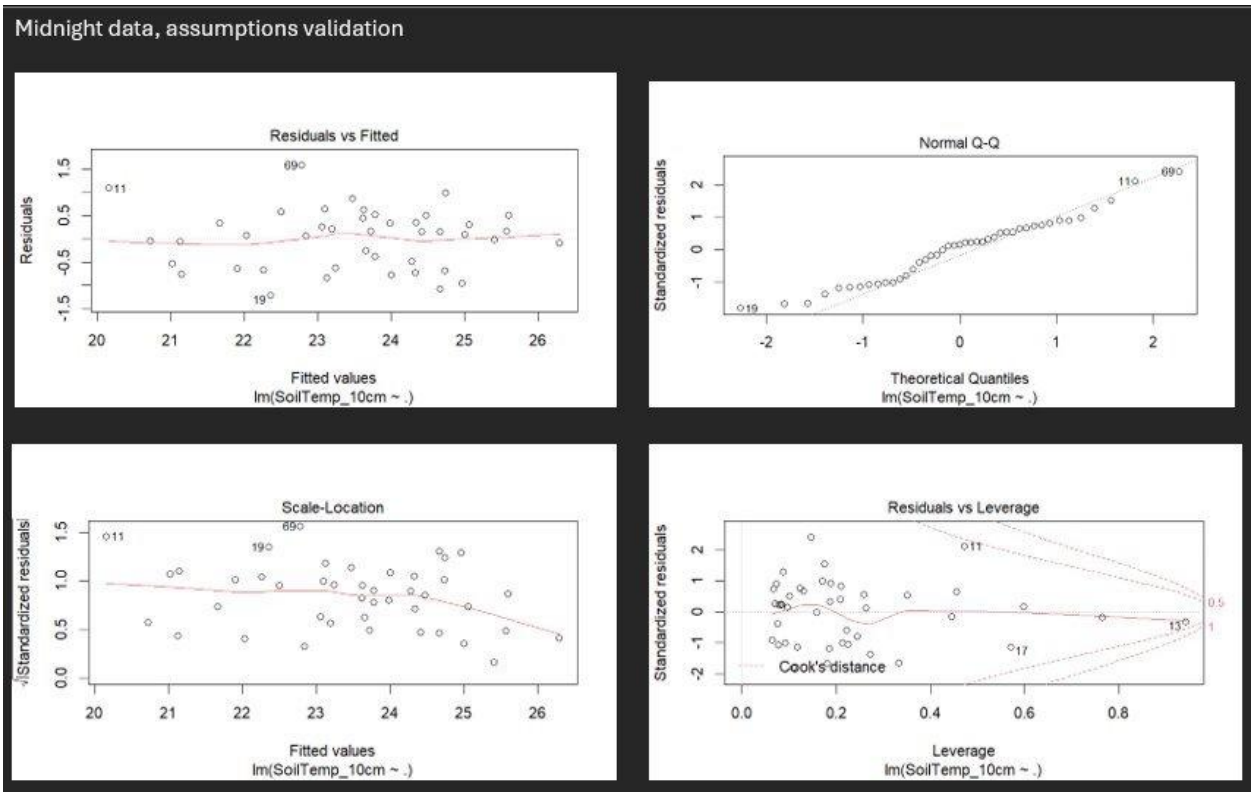
Noticeably, none of the variables are statistically significant (all their p-values exceed the referential 5%). We then fail to reject the null hypothesis and conclude that at 5% of significance, there is not enough evidence from the sampled data that the coefficients are equal to zero.

Nevertheless, the overall model is significant given the F statistic p-value reaching nearly zero. This leads us to believe that either there are very many predictors in the model and that each makes their contribution to making the model be significant, or these predictors are strongly correlated amongst each other.

As for the first suspicion, we notice that 83.97 percent of variability in SoilTemp_10cm is explained by all predictors in the regression model when appreciating the R-squared result. This is a strong number that may suggest overfitting of the model. As for the second suspicion, we measured multicollinearity by calculating the Variance Inflation Factor. These values excessively exceed the referential value of 10:

```
vif(mlr_midnight)

AirTemp_6ft_CP      RH_6ft_CP      DP_6ft_CP AirTemp_6ft_T1      RH_6ft_T1
     185676.71      207425.11      552478.20       78341.18      117143.14
     DP_6ft_T1 AirTemp_6ft_T2      RH_6ft_T2      DP_6ft_T2
     407801.71       80114.25       95267.15      107452.60
```

Therefore, we confirm both suspicions and draw the conclusion, among others, that the multiple linear regression may not be the best statistical method to understand the impact (or significant impact) of the predictors to the soil temperature. As a result, we continue our analysis with other statistical methods. But before that, we validate the assumptions of this model...



Linearity: The Residuals vs Fitted graph shows randomness of points with values centered around zero and no trend patterns found in Graph. This indicated the Linearity assumption is valid.
Normality: The normal Q-Q plot shows that the points fall along the straight line and there is three or four points are somewhat far from the line. Given that this is a robust test, we found the Normality assumption to not be severely violated.
Homoscedasticity: It looks like residuals have constant variance against the fitted values. This shows that the Homoscedasticity assumption is validated.
Outliers: no strong points are found to be highly influential. As a result, we consider the Outliers assumption to not be violated.

Regularized Regression (Ridge and Lasso Regression)

The main outputs for each Ridge and Lasso Regression are respectively shown on the 2 lists below:

	s0
(Intercept)	3.9678081908
AirTemp_6ft_CP	0.0484070870
RH_6ft_CP	-0.0002257969
DP_6ft_CP	0.0102232146
AirTemp_6ft_T1	0.0444036800
RH_6ft_T1	-0.0006972659
DP_6ft_T1	0.0073926425
AirTemp_6ft_T2	0.0478617506
RH_6ft_T2	-0.0018140431
DP_6ft_T2	0.0084853299
SoilTemp_10cm	0.6830128056

	s0
(Intercept)	11.53829892
AirTemp_6ft_CP	.
RH_6ft_CP	.
DP_6ft_CP	0.02283532
AirTemp_6ft_T1	.
RH_6ft_T1	.
DP_6ft_T1	.
AirTemp_6ft_T2	0.50378081
RH_6ft_T2	.
DP_6ft_T2	.

For Ridge, the best lambda (our tuning parameter to control the suggested overfitting mentioned above after applying cross-validation) was 0.1553 with an R-squared of 0.9803. All predictors approximate to zero simplifying the model greatly and making us suspect it needs to be reduced. For this purpose, Lasso regression is applied.

For Lasso, the best lambda was 0.0647 with an R-squared of 0.8056. After the consequent feature selection applied, we find that the Dew Point at 6 feet on the post (DP_6ft_CP) and the Air Temperature at 6 feet on Tree 2 are the two (2) predictors that most matter. All the coefficients of the other variables were set to zero.

If the intention is to choose the simplest model, then Lasso regression yields better results than Ridge. But if we aim to have a model that best describes its variability, Ridge is a better model.

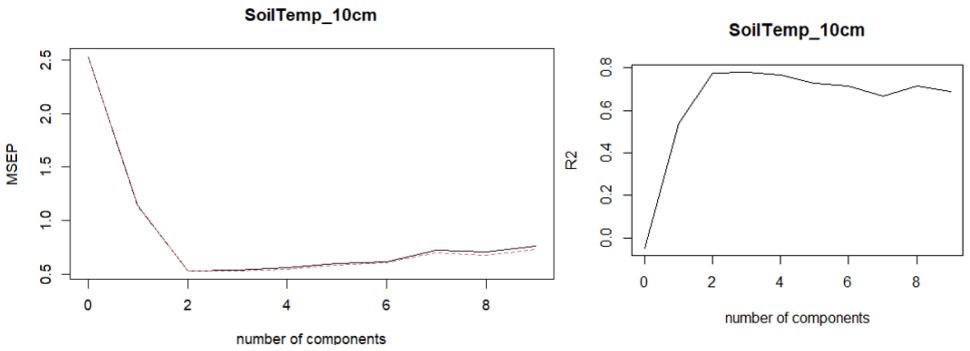
Principal Component Analysis

Principal Component Regression

To address the biggest problem in our model (multicollinearity), PCR is applied to transform the highly correlated predictors into uncorrelated variables called principal components.

We then chose 2 to be the number of principal components considering the small marginal gain in explained variance if we had 3 components. We note that with only 2 components we obtain:

- Mean Squared Error of Prediction (MSEP): 0.7326
- % Variance: 99.99%



Model 3

Model 3, as indicated on the Introduction, replicates Model 2 with the difference that, instead of covering the midnight hours, the midday hours are covered (from 12pm – 1pm). Again, the values of the predictors and the dependable variable were averaged while preserving the same meaning. The idea behind building Model 3 is to enable a comparison of the expected "peak" (or highest point) of the soil temperature against the "valley" (or lowest point) of the soil temperature during the day.

Given the above, many of the explanatory details expounded on Model 2 will be omitted while keeping the focus on the results and the actual discussion.

Multiple Linear Regression

The Model 3 is described by the following equation:
SoilTemp_10cm = 59.383 - 5.415(AirTemp_6ft_CP) - 1.408(RH_6ft_CP) + 3.303(DP_6ft_CP) - 7.904(AirTemp_6ft_T1) -1.166(RH_6ft_T1) + 11.919(DP_6ft_T1) + 11.197(AirTemp_6ft_T2) + 2.103(RH_6ft_T2) - 12.630(DP_6ft_T2).

That is, during noon (from 12am to 1am), the average temperature of the soil at 10 cm of depth expressed in Celsius can be described by the averaged Air Temperature, the Relative Humidity, and the Dew Point at 6 feet height on the post, at 6 feet height of the Tree 1, and at the same height on Tree 2, respectively.

The specific interpretation of the coefficients of the model is as follows... each coefficient represents the change in soil temperature associated with a one-unit change in the predictor variable, holding other variables constant:

- The soil temperature would start at 59.383 degrees Celsius if none of the predictors were present. This hypothetical average is clearly higher than the result in the case of midnight.
- AirTemp_6ft_CP: A one-unit increase in the average air temperature at 6 feet in location CP (the post) decreases soil temperature by about 5.415 units... a contrary effect from the case of midnight, obviously.
- RH_6ft_CP: A one-unit increase in the average relative humidity at 6 feet in location CP (the post) decreases soil temperature by about 1.408.
- DP_6ft_CP: A one-unit increase in the average dew point at 6 feet in location CP (the post) increases soil temperature by about 3.303 units.
- AirTemp_6ft_T1: A one-unit increase in the average air temperature at 6 feet in location T1 (tree 1) decreases soil temperature by about 7.904 units.
- RH_6ft_T1: A one-unit increase in the average relative humidity at 6 feet in location T1 (tree 1) decreases soil temperature by about 1.166 units.
- DP_6f_T1: A one-unit increase in the average dew point at 6 feet in location T1 (tree 1) increases soil temperature by about 11.919 units.
- AirTemp_6ft_T2: A one-unit increase in the average air temperature at 6 feet in location T1 (tree 1) increases soil temperature by about 11.197 units.
- RH_6ft_T2: A one-unit increase in the average relative humidity at 6 feet in location T2 (tree 2) increases soil temperature by about 2.103 units.
- DP_6ft_T2: A one-unit increase in the average dew point at 6 feet in location T2 (tree 2) decreases by about 12.630 units.

Compared to Model 2, the coefficients in Model 3 behave in the opposite direction. That is, the relative humidity, the air temperature, and the dew point cause a contrary effect during at midnight versus the midday.

```
Call:
lm(formula = SoilTemp_10cm ~ ., data = midday_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5995 -0.3978 -0.1559  0.2965  2.7677

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     59.383     53.780   1.104    0.278
AirTemp_6ft_CP  -5.415      8.883  -0.610    0.546
RH_6ft_CP       -1.408      1.955  -0.720    0.476
DP_6ft_CP        3.303     10.377   0.318    0.752
AirTemp_6ft_T1  -7.904     11.501  -0.687    0.497
RH_6ft_T1       -1.166      2.322  -0.502    0.619
DP_6ft_T1       11.919     12.727   0.936    0.356
AirTemp_6ft_T2  11.197     12.402   0.903    0.373
RH_6ft_T2        2.103      2.488   0.845    0.404
DP_6ft_T2      -12.630     13.108  -0.964    0.342

Residual standard error: 0.8855 on 33 degrees of freedom
Multiple R-squared:  0.8044,    Adjusted R-squared:  0.7511
F-statistic: 15.08 on 9 and 33 DF,  p-value: 2.423e-09
```

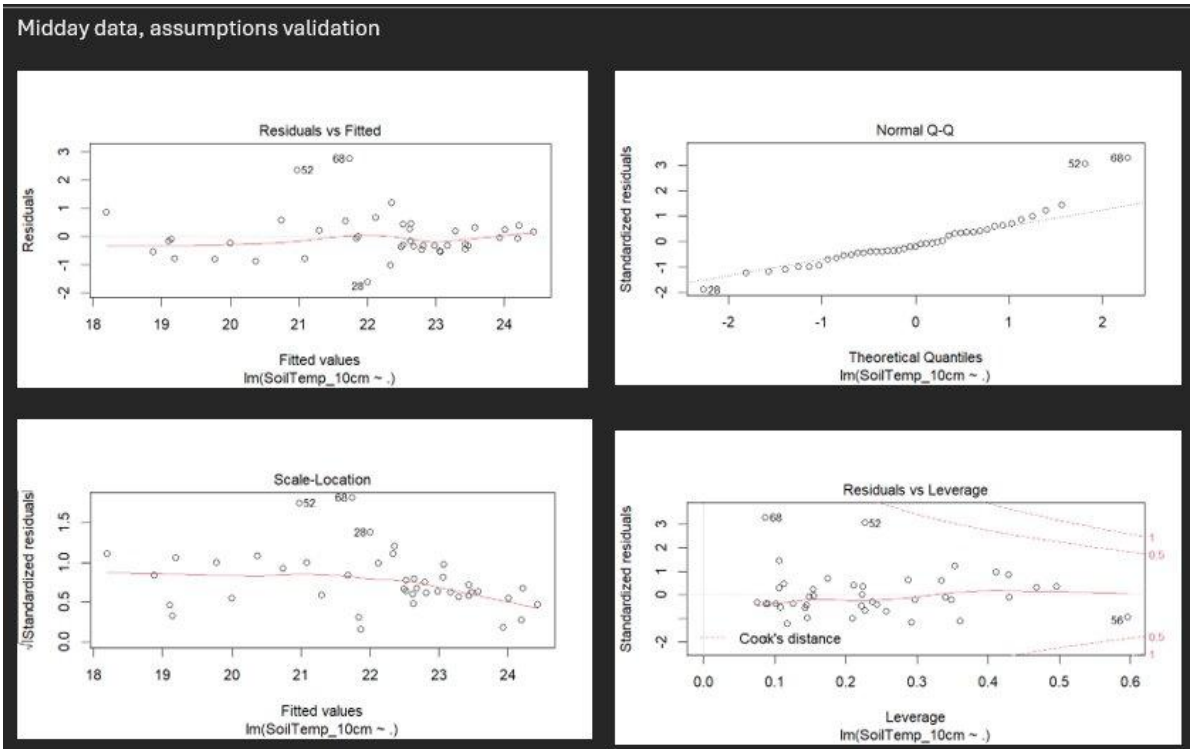
Like in Model 2, the degree of variation is highly explained by the model when looking at the R-squared results: 83.96% versus 80.44%. This may again suggest an overfitting like Model 2.

Notice that the same effect of coefficients significance and the overall model significance persist in Model 3. The same interpretation for Model 2 then holds in Model 3. In summary, we see a model statistically significant with all variables being not statistically significant. Given the two (2) suspicions expounded in Model 2 (all predictors making their own individual contribution to the model and a strong multicollinearity amongst them) are again confirmed. Below, the results of the Variance Inflation Factor for Model 3:

```
vif(mlr_midday)

AirTemp_6ft_CP      RH_6ft_CP      DP_6ft_CP AirTemp_6ft_T1      RH_6ft_T1
      52238.54       12145.45       80661.12       78524.11       16164.42
     DP_6ft_T1 AirTemp_6ft_T2      RH_6ft_T2      DP_6ft_T2
     118479.24       99499.54       18767.06      128821.11
```

Our approach keeps consistent with the statistical analysis. As done with Model 2, we now look at other statistical methods given the issues described above in the multiple linear regression analysis. But first, the validation of the assumptions of the model…

The Linearity, Normality, Homoscedasticity and Outliers assumptions are considered valid in Model 3 for the same reasons as expounded in Model 2. Below, the results in graphs.



Regularized Regression (Ridge and Lasso Regression)

The main outputs for each Ridge and Lasso Regression are respectively shown on the 2 lists below:

	s0
(Intercept)	4.394643196
AirTemp_6ft_CP	0.030898771
RH_6ft_CP	-0.006878319
DP_6ft_CP	0.021193634
AirTemp_6ft_T1	0.017943462
RH_6ft_T1	0.004562723
DP_6ft_T1	0.018196997
AirTemp_6ft_T2	0.022666247
RH_6ft_T2	-0.002079504
DP_6ft_T2	0.019991623
SoilTemp_10cm	0.702891585

	s0
(Intercept)	18.9101456
AirTemp_6ft_CP	.
RH_6ft_CP	-0.1960218
DP_6ft_CP	.
AirTemp_6ft_T1	.
RH_6ft_T1	0.1419559
DP_6ft_T1	0.4291424
AirTemp_6ft_T2	.
RH_6ft_T2	.
DP_6ft_T2	.

For Ridge, the best lambda (our tuning parameter to control the suggested overfitting mentioned above after applying cross-validation) was 0.1754 with an R-squared of 0.9805. These results are quite similar to those obtain for Model 2 (0.1553 and 0.9803, respectively). Again, all predictors approximate to zero simplifying the model greatly and making us suspect that it needs to be reduced. For this purpose, Lasso regression is applied.

For Lasso, the best lambda was 0.0511 with an R-squared of 0.9992. After the consequent feature selection applied, we find that the relative humidity both at 6 feet on the post and the tree 1 as well as the dew point at 6 feet on tree 1 are the three (3) variables that matter the most.

If the intention is to choose the simplest model, then Lasso regression yields better results than Ridge. But if we aim to have a model that best describes its variability, Ridge is a better model.

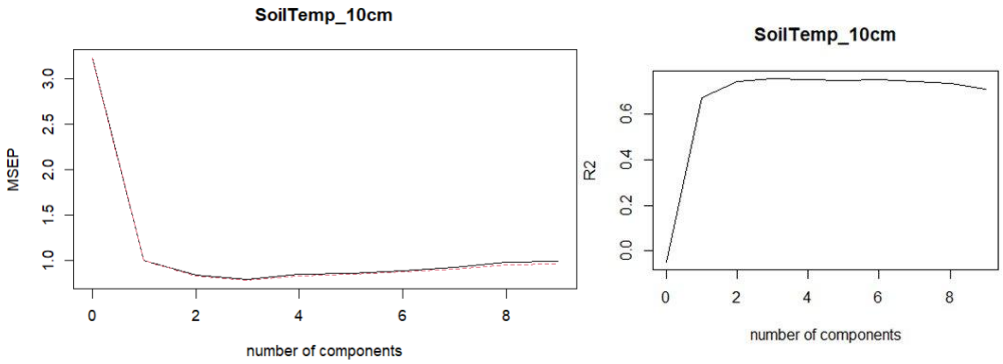
Principal Component Analysis

Principal Component Regression

To address the biggest problem in our model (multicollinearity), PCR is applied to transform the highly correlated predictors into uncorrelated variables called principal components.

We then chose 2 to be the number of principal components considering the small marginal gain in explained variance if we had 3 components. We note that with only 2 components we obtain:

- Mean Squared Error of Prediction (MSEP): 0.8657
- % Variance: 99.94%



In comparison to Model 2, Model 3 yields a higher MSE (86.57% versus 73.26%) with also an incredibly strong percent of variance.

Model 4
Multiple Linear Regression

This output from the linear regression analysis describes how soil moisture at 15 cm depth (SoilMoist_15cm_Average) relates to several climate variables at different locations and heights. The resulted model will look like:

Soil Moisture (15cm) = 0.3422 + RH_6ft_CP_Average(0.0247) + DP_6ft_CP_Average(0.395) + RH_6ft_T1_Average(0.0011) - DP_6ft_T1_Average(0.409) - RH_6ft_T2_Average(0.0271) + DP_6ft_T2_Average(0.01516)

Model Summary:

Intercept: The estimated intercept of 0.3422 suggests that when all other variables are set to zero, the average soil moisture at 15 cm is expected to be approximately 0.3422 units. This value is statistically significant with a p-value of 0.00427, indicating that it is unlikely to be zero by chance.

Coefficients: Each coefficient represents the change in soil moisture associated with a one-unit change in the predictor variable, holding other variables constant:
RH_6ft_CP_Average: A one-unit increase in the average relative humidity at 6 feet in location CP increases soil moisture by about 0.0248 units (significant at $p < 0.05$).
DP_6ft_CP_Average: A one-unit increase in the average dew point at 6 feet in location CP increases soil moisture by about 0.3952 units (near-significant at $p < 0.1$).
DP_6ft_T1_Average: A one-unit increase in the average dew point at 6 feet in location T1 decreases soil moisture by about 0.4090 units (significant at $p < 0.05$).

Other variables like RH at T1 and T2 and DP at T2 showed smaller and statistically non-significant changes.

Residuals:
The residuals of the model (differences between observed and predicted soil moisture values) range from -0.08245 to +0.18499. The median near zero (-0.01926) indicates that the model does not systematically over- or under-predict across the data set.

Model Fit:
Residual Standard Error (RSE): The RSE of 0.05639 suggests that the typical prediction error is about 0.05639 units of soil moisture.

Multiple R-squared: The value of 0.2943 indicates that approximately 29.43% of the variability in soil moisture is explained by the model. This is a moderate amount, suggesting other factors not included in the model also influence soil moisture.

Adjusted R-squared: The adjusted R-squared of 0.1767 is lower than the multiple R-squared, reflecting the penalty for the number of predictors in the model and suggesting that some predictors might not be contributing effectively to the model.

F-statistic: An F-statistic of 2.502 with a p-value of 0.03977 shows that the model is statistically significant overall, indicating that there is a relationship between the predictors and the response variable.

Implications:
The model shows some significant relationships but also highlights that some predictors are not significantly influencing soil moisture, and overall model fit could be improved. This could lead to considerations of additional data collection, exploring non-linear relationships, or including other potentially relevant variables.

```
##Linear Regression
lm = lm(SoilMoist_15cm_Average~.,data = data )
summary(lm)


Call:
lm(formula = SoilMoist_15cm_Average ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.08245 -0.03360 -0.01926  0.02639  0.18499

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.342205   0.112182   3.050  0.00427 **
RH_6ft_CP_Average  0.024763   0.012021   2.060  0.04669 *
DP_6ft_CP_Average  0.395172   0.215286   1.836  0.07469 .
RH_6ft_T1_Average  0.001103   0.013767   0.080  0.93661
DP_6ft_T1_Average -0.409008   0.200559  -2.039  0.04881 *
RH_6ft_T2_Average -0.027176   0.014898  -1.824  0.07644 .
DP_6ft_T2_Average  0.015167   0.191972   0.079  0.93747
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05639 on 36 degrees of freedom
Multiple R-squared:  0.2943,    Adjusted R-squared:  0.1767
F-statistic: 2.502 on 6 and 36 DF,  p-value: 0.03977
```

Assumptions

Residuals vs. Fitted: This plot is used to detect non-linearity, unequal error variances, and outliers. Ideally, the residuals should be randomly scattered around the horizontal line (zero), indicating a good fit. In your plot, the residuals do not appear to display any systematic pattern, which is good, but there might be a slight curve, suggesting potential non-linearity in the relationship between predictors and the response variable.

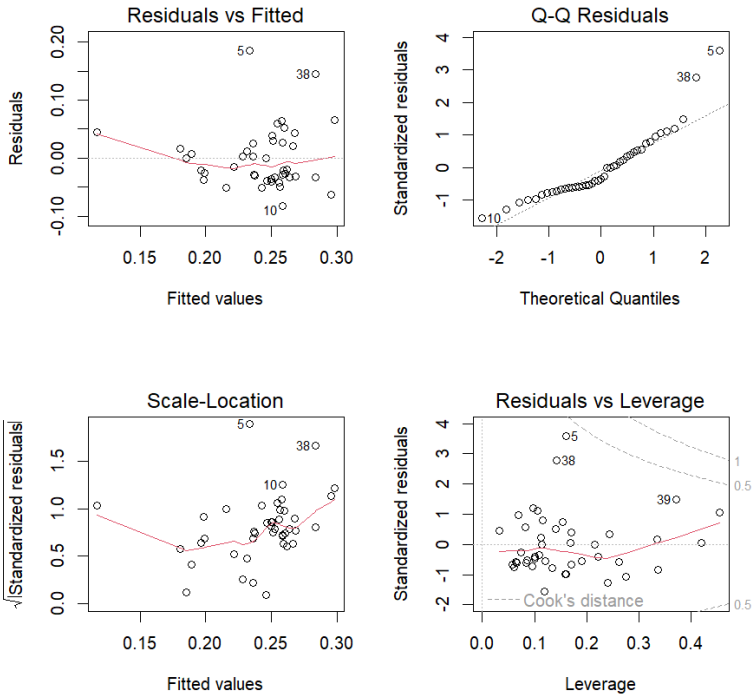
Q-Q (Quantile-Quantile) Plot: This is used to check the normality of residuals. The points should fall approximately along the reference line if the residuals are normally distributed. Your plot shows that the residuals roughly follow the line, but with some deviations at the tails. This could indicate slight violations of normality.

Scale-Location (Spread-Location) Plot: It is used to check the homoscedasticity (equal variance of residuals). If the residuals are equally spread across all levels of the fitted values, the assumption is met. The red line should be horizontal and flat if homoscedasticity holds. The plot shows some signs of non-constant variance, as the red line has a slight "smile" pattern.

Residuals vs. Leverage: This helps identify influential observations that could have a significant impact on the regression line. Points that stand out far to the right or have a high Cook's distance (outside the dashed Cook's distance lines) are of particular interest. In this plot, while there are a

few observations with higher leverage, they do not seem to have high Cook's distance values, indicating they may not be unduly influencing the model.

Overall, the diagnostic plots suggest that the linear regression model fits reasonably well, although there might be some concerns with non-linearity and non-constant variance of residuals.



Mutlicollinearity

As all the values from the VIF function is greater than 10 which shows all the variables have multicollinearity between them and they are not independent of each other and the presence of such multicollinearity can distort the regression coefficients and their standard errors, leading to unreliable p-values and potentially incorrect conclusions.

```
> vif(lm)
RH_6ft_CP_Average DP_6ft_CP_Average RH_6ft_T1_Average DP_6ft_T1_Average
        113.2413         8560.5707          140.1512         7254.2510
RH_6ft_T2_Average DP_6ft_T2_Average
        165.9347         6812.9533
```

Regularized Regression

Ridge Model

These results are affected by the Ridge regression process, which shrinks coefficients as a way to handle multicollinearity and improve model generalization. The coefficients for the relative humidity variables are very close to zero, suggesting that, after accounting for multicollinearity, these predictors have a minimal impact on the response variable in the presence of other predictors.

```
best_model = glmnet(x, y, alpha = 0, lambda = best_lambda)
coef(best_model)

7 x 1 sparse Matrix of class "dgCMatrix"
                              s0
(Intercept)        2.090411e-01
RH_6ft_CP_Average -3.245238e-05
DP_6ft_CP_Average  7.294607e-04
RH_6ft_T1_Average  3.726528e-05
DP_6ft_T1_Average  7.097600e-04
RH_6ft_T2_Average -6.760653e-05
DP_6ft_T2_Average  7.193943e-04
```

Lasso Model

Lasso regression is known for its ability to perform variable selection by shrinking coefficients of less relevant predictors to zero. The results indicate that, aside from the dew point at a 6-foot height at location CP and the model intercept, all other variables were deemed non-significant by the lasso model and thus their coefficients were reduced to zero. This suggests that the dew point at this specific location, along with the baseline soil moisture level indicated by the intercept, accounts for approximately 30% of the variance in the model's predictive ability.

```
best_modell <- glmnet(x, y, alpha = 1, lambda = best_lambdal)
coef(best_modell)

7 x 1 sparse Matrix of class "dgCMatrix"
                             s0
(Intercept)       0.210985242
RH_6ft_CP_Average .
DP_6ft_CP_Average 0.001753191
RH_6ft_T1_Average .
DP_6ft_T1_Average .
RH_6ft_T2_Average .
DP_6ft_T2_Average .
```

Principal Component Analysis

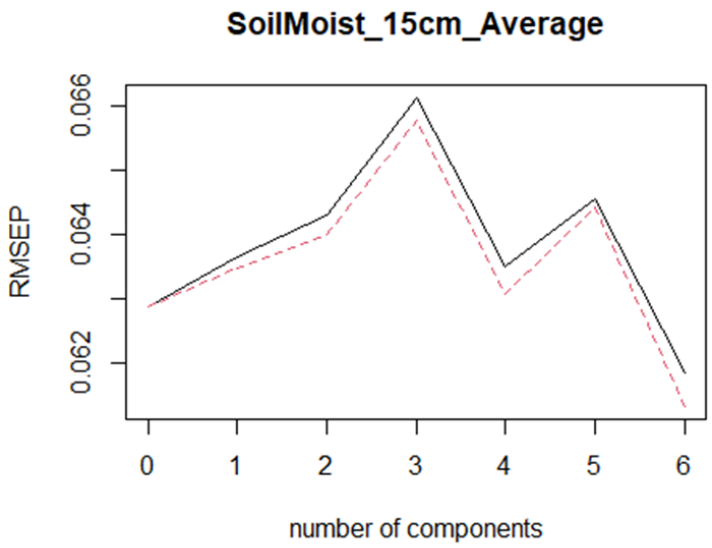
Principal Component Regression

The results of the principal component analysis suggest that four components account for a substantial portion of the variance in relation to the response variable. Therefore, we will select four variables for our model.

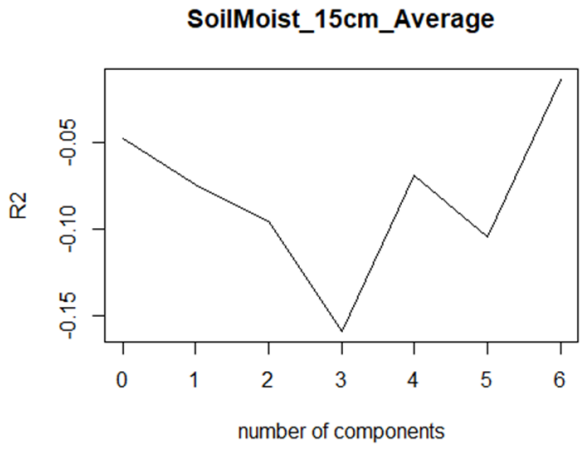
```r
pca = pcr(SoilMoist_15cm_Average~., data=data, scale=TRUE, validation="CV")
summary(pca)
```

```
Data:   X dimension: 43 6
    Y dimension: 43 1
Fit method: svdpc
Number of components considered: 6

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV         0.06289  0.06367  0.06430  0.06614  0.06352  0.06455  0.06185
adjCV      0.06289  0.06349  0.06401  0.06577  0.06308  0.06442  0.06133

TRAINING: % variance explained
                        1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
X                        71.413   99.733   99.928   100.00   100.00   100.00
SoilMoist_15cm_Average    3.613    8.758    9.957    20.09    21.85    29.43
```

```r
validationplot(pca)
```

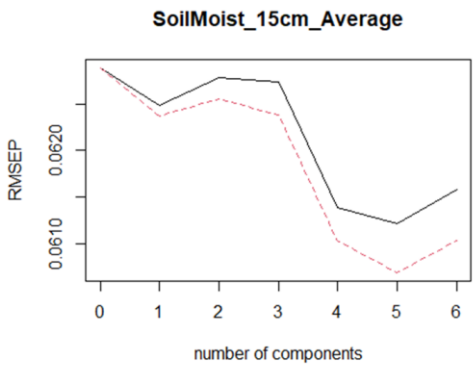


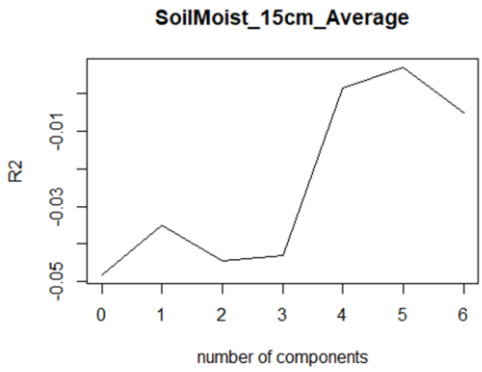

```r
validationplot(pca, val.type = "R2")
```



Partial Least Square

```
plsModel=plsr(SoilMoist_15cm_Average~., data=data, scale=TRUE,
validation="CV")
summary(plsModel)

Data:   X dimension: 43 6
    Y dimension: 43 1
Fit method: kernelpls
Number of components considered: 6

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV         0.06289  0.06249  0.06278  0.06274  0.06138  0.06121  0.06158
adjCV      0.06289  0.06237  0.06256  0.06238  0.06103  0.06069  0.06103

TRAINING: % variance explained
                        1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
X                        63.530   99.731    99.86   100.00   100.00   100.00
SoilMoist_15cm_Average    7.261    8.851    18.33    20.59    29.31    29.43

validationplot(plsModel)
```





validationplot(plsModel, val.type = "R2")

CONCLUSIONS

- Model 1 indicates that soil temperature is significantly influenced by the parameters of dew point, humidity, and air temperature, with an R-squared value of 0.82. This high R-squared value suggests that the model effectively explains the variation in soil temperature.
- Due to multicollinearity among the independent variables, ridge and lasso regression techniques were employed. Lasso led to the elimination of many variables, ultimately narrowing down the predictors to one or only a few.
- The average soil temperature is higher during the night and cooler at noon. The soil takes half of a day to reach the "peak" and "valley" values in the 24-hour cycle as it cools off until midday and starts warming after that. The change in temperature cycles contrary to the temperature (for instance, when the air temperature is hotter, the soil temperature is still cool).
- The multiple regression models seem unfit for this climate data whereas, depending on the intended goal, other regression models provide a better explanation. Such is the case for Ridge and Lasso. The former explains the variability better and the latter explains the model simpler (with fewer variables).
- We may still choose to apply time series analysis to understand other hidden factors, such as cycles, seasonality, trends, and more patterns. This would entail at least extending the sample size.
- Model 4 fails to capture the influence of independent variables such as dew point and humidity on soil moisture, as indicated by a low R-squared value of 0.29. This suggests that other factors, such as rainfall, might better explain the variations in soil moisture.